DOCUMENT RESUME

ED 411 283 TM 027 288

AUTHOR	Barcikowski, Robert S.; Elliott, Ronald S.
TITLE	Pairwise Multiple Comparisons in Single Group Repeated Measures Analysis.
PUB DATE	1997-03-00
NOTE	30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Comparative Analysis; *Educational Research; Monte Carlo Methods; *Research Design; Simulation
IDENTIFIERS	Bonferroni Procedure; Pairwise Preference Data; *Repeated Measures Design; *Single Group Design; T Test; Variance (Statistical)

ABSTRACT

Research was conducted to provide educational researchers with a choice of pairwise multiple comparison procedures (P-MCPs) to use with single group repeated measures designs. The following were studied through two Monte Carlo (MC) simulations: (1) The T procedure of J. W. Tukey (1953); (2) a modification of Tukey's T (G. Keppel, 1973); (3) the Dunn-Bonferroni procedure (DB); (4) the sequentially rejective Bonferroni procedure (J. P. Shaffer, 1986) (SB); (5) Hayter's modification of Fisher's Least Significant Difference Test (A. J. Hayter, 1986) (FH); (6) a modified range procedure combining others (SRW); (7) a multiple range procedure based on Ryan-Welsch critical values (MRW) (T. A. Ryan and R. E. Welsch); (8) E. Peritz's (1970) procedure (P); and (9) Welsch's step-up procedure (W). The first MC study was exploratory and was based on variance-covariance matrices that were created so as to conform to different sphericity values. Power in this study was examined for a fixed set of mean differences. The second MC study, based on the results of the first, used variance-covariance matrices found in 100 real repeated measures data sets. Based on study results, the stepwise tests SB, FH, SRW, MRW, and P and the T and W are not recommended, but the DB procedure is recommended for use with single group repeated measures data. (Contains 5 tables, 10 figures, and 34 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED 411 283

Pairwise Multiple Comparisons In Single Group Repeated Measures Analysis

Robert S. Barcikowski
Ohio University
and
National Institute For Educational Development, Okahandja, Namibia

Ronald S. Elliott
Ohio University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Ronald Elliott

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM027288

BEST COPY AVAILABLE

A paper presented at the annual meeting of the American Educational Research Association, Chicago, March, 1997.

Pairwise Multiple Comparisons In Single Group Repeated Measures Analysis

Introduction

The main purpose of this research was to provide educational researchers with a choice of pairwise multiple comparison procedures (P-MCPs) to use with single group repeated measures data. This was done through two Monte Carlo (MC) studies. The first MC study[1] was exploratory and was based on variance-covariance matrices that were created so as to conform to different sphericity values. Power in this study was examined for a fixed set of mean differences. The second MC study was based on the results of the first study, and used variance-covariance matrices found in one-hundred real repeated measures data sets. Power in the second study was examined based on the mean differences found in these real data.

Study 1

Objectives

The first objective in study 1 was to examine P-MCPs that have been shown to control different types of Type 2 error and Type 1 familywise error under both no violations and violations of assumptions in other designs. A second objective, was to recommend one or more of the P-MCPs to educational researchers based on ease of use. This study expanded the previous work done in this area (e.g., Maxwell (1980), Boik (1981), Alberton and Hochberg (1984), Keselman, Keselman and Shaffer (1991), Keselman (1994), Keselman and Lix (1995)) by:

(a) using Bradley's(1978) stringent level of robustness to examine the P-MCPs empirical rate of Type I error ($\hat{\alpha}$) as compared with the nominal familywise level of significance (α);
(b) expanding the range of sphericity (as measured by ε) considered to more realistically cover those values found in practice (Green and Barcikowski, 1992);
(c) comparing per-pair power among the P-MCPs by finding the number of units (n's) necessary to reach per-pair power of .80.

[1] The first study was presented as a paper at the annual meeting of the Mid-Western Educational Research Association, Chicago, October, 1996.

Perspectives

P-MCPs Studied

A great deal of work has been done recently in the development of new and competing P-MCPs (Seaman, Levin, and Serlin, 1991). Many of these new P-MCPs have been adapted for use in split-plot repeated measures designs in papers written by the Keselmans and their colleagues (Keselman, Keselman and Shaffer (1991), Keselman Carriere and Lix (1993), Keselman (1994), Keselman and Lix (1995)). In this paper the following P-MCPs, described in detail by Maxwell (1980), Keselman (1994), and Keselman and Lix (1995) were examined for use with single group repeated measures data: 1) Tukey's T procedure (also known as the Studentized range procedure) (Tukey, 1953), 2) A modification of Tukey's T suggested by Keppel (1973) and studied by Maxwell (1980), 3) Dunn-Bonferroni controlled t-tests (DB), 4) Shaffer's (1986) sequentially rejective Bonferroni procedure (SB), 5) Hayter's (1986) two-stage modification of Fisher's Least Significant Difference test (FH), 6) A modified range procedure that combines the work of Shaffer(1979, 1986), Ryan(1960) and Welsch (1977) (SRW), 7) A multiple range procedure based on Ryan-Welsch critical values (MRW), 8) Peritz's (1970) procedure (P), and 9) Welsch's (1977) step-up procedure (W).

These P-MCPs were selected for study because they were found to be at least partially successful in controlling different types of Type 2 error and Type 1 familywise error in previous studies. The first three procedures were used by Maxwell (1980) in his study of this problem, and procedures 4 through 8 were found by Keselman and Lix (1995) to be robust to violations of normality, multisample sphericity and heterogeneity of variance-covariance matrices with unequal cell sizes in split-plot designs using Bradley's liberal criterion. Keselman and Lix (1995) examined procedures 4 through 8 using the Welch-James-Johansen (WJJ) overall multivariate test (Johansen, 1980) with Satterthwaite (1941) adjusted degrees of freedom (SDF) as described by Keselman, Keselman and Shaffer (1991). They also modified the range procedures (SRW, MRW, P) by using a process described by Duncan (1957). Keselman (1994) recommended the Welsh step-up procedure with SDF degrees of freedom for use with split-plot repeated measures designs over twenty-seven other methods that he studied. Therefore, the first three procedures are generally familiar to most educational researchers and they provided check points with Maxwell's study. The second six procedures were found to be effective under more severe violations of assumptions, and were expected to perform well in this study of a simpler design.

The T, K, DB, and W P-MCPs were studied without an overall test. The T, K, DB P-MCPs are called *simultaneous* procedures because they use a single critical value to test all pairwise differences. The SB, FH, SRW, MRW, P and W are referred to as *stepwise* or *sequential* procedures because they test stages of hypotheses in a stepwise fashion, usually using a different critical value at each stage. SB, FH, SRW, MRW, and P were to be examined after first being

preceded by the WJJ test. The FH procedure was to be studied after being preceded by Keppel's q-statistic based on the Studentized-range. The SRW, MRW, and P range procedures were to be conducted with the modification described by Duncan (1957).

Background Equations

The P-MCPs examined in this study may be better understood through the following set of equations. In the following equations we are comparing pairs of means (i,j) from a set of J means where i, j = 1, 2, ..., J and $i \neq j$. Then, S^2 is the mean square error (i.e., the mean square within, or residual) of the analysis of variance considered, and S_i^2 and S_j^2 are the variances of treatments or measures i and j, with sample sizes n_i and n_j, respectively. When all treatments or measures have an equal number of units, the treatment or measure sample size is denoted by n. The general form of these equations is found in Equation 1.

Equation 1: General Form.

$$TS_{ij} \geq CV_{ij,\alpha,v} * Con \qquad (1)$$

The term TS_{ij} is the calculated test statistic in the form of a *t* statistic for various situations, and the term $CV_{ij,\alpha,v}$ is a critical value with familywise error of α and error degrees of freedom v. The term Con is a constant which allows the equation to be valid. When the calculated test statistic TS_{ij} is greater than or equal to $CV_{ij,\alpha,v}$ times Con, mean i is said to differ significantly from mean j.

Equation 2: Equal n, Homogeneous Variances.

$$TS_{ij} = (\bar{Y}_i - \bar{Y}_j) / (S^2 / n)^{1/2} \geq CV_{ij,\alpha,v} * CON \qquad (2)$$

The typical example for this equation is Tukey's HSD used to compare all pairs of means in a one-way ANOVA with J treatments. Then, $CV_{ij,\alpha,v}$ is the Studentized Range Statistic and Con = 1.0. For example, in a one-way ANOVA with J = 5, n = 9 units (e.g., subjects) per treatment, and $\alpha = .05$, we have that $CV_{ij,.05,40} = q_{\alpha,j,v} = q_{.05,5,40} = 4.04$ for all paired comparisons.

Equation 3: Unequal n, Homogeneous Variances.

$$TS_{ij} = (\bar{Y}_i - \bar{Y}_j) / (S^2 / n_i + S^2 / n_j)^{1/2} \geq CV_{ij,\alpha,v} * CON \qquad (3)$$

Equation 4: Unequal n, Heterogeneous Variances.

$$TS_{ij} = (\bar{Y}_i - \bar{Y}_j) / (S_i^2 / n_i + S_j^2 / n_j)^{1/2} \geq CV_{ij,\alpha,v} * CON \qquad (4)$$

Equation 5: Equal n, Heterogeneous Variances, correlated measures.

$$TS_{ij} = (\bar{Y}_i - \bar{Y}_j) / ((S_i^2 + S_j^2 - 2S_{ij}) / n)^{1/2} \geq CV_{ij,\alpha,v} * CON \qquad (5)$$

Where S_{ij} is the covariance between measures i and j and for single group repeated measures designs v is usually equal to n-1.

Equation 5 may be used to illustrate all of the P-MCPs considered in this study, except the T procedure which uses Equation 2. This can be done with the assistance of Table 1 which provides information on the test statistics and how their levels of significance and "steps between means" degrees of freedom are determined in order to control familywise error rate. *Familywise error* (FWE) is the probability of making at least one Type I error when testing a family of hypotheses.

An example of where Equation 5 might be used is in a single group repeated measures analysis with J = 7 measures on n = 25 subjects. Maxwell (1980) recommended the Dunn-Bonferroni approach to determine which pairs of means differed. Using the Dunn-Bonferroni approach, and the aid of Equation 5 and Table 1, we have that $CV_{ij,\alpha,v}$ is student's t-statistic with $\alpha' = 2\alpha/(J*(J-1)) =$ .00238 and v = n-1 = 24 degrees of freedom. Then, $CV_{ij,.05,24} = t_{.00238,24} =$ 3.396 and Con = 1.0 for all paired comparisons.

Method

Design characteristics. The complexity and number of conditions to be compared necessitated a Monte Carlo study. In order to investigate the Type 1 and Type 2 error rates the following characteristics of the single group design were manipulated: (1) the number of repeated measures (J = 3, 4, 5, 6, 8, 10), (2) the value of sphericity (at J = 3 ε = .51, .75, and 1.0; for each other J four values of ε were examined, ε = .50, .75, and 1.0 plus a value near the minimum for ε,

i.e., for J = 4, ε = .40; J = 5, ε = .30; J = 6, ε = .30; J = 8, ε = .20; J = 10, ε = .20;), and (3) the shape of the population (normal, nonnormal with skewness = 1.75, and kurtosis = 3.75). The variance-covariance matrices for each value of sphericity were generated using an algorithm developed by Cornell, Young and Bratcher (1990). Probabilities and upper quantiles for the Studentized Range statistic (q) were computed through an algorithm developed by Lund and Lund (1983). The number of repeated measures and the values of sphericity were based on a study by Green and Barcikowski (1992) and the shape of the nonnormal distribution was close to that chosen by Keselman (1994) (skewness = 1.633, and kurtosis = 4.0), based on an investigation by Micceri(1989).

Data generation. A FORTRAN program was used to generate the repeated measures normal data following procedures described by Barcikowski (1980). Each covariance matrix (C) was factored into upper and lower triangular matrices L and L' using the Cholesky (square root factoring) decomposition of R, i.e., R = L L'. Repeated measures for a unit (e.g., subject) were arrived at using a procedure described by Collier, Baker, Mandeville and Hayes (1967, pp. 343-344) where a vector of J scores, $\underline{z}$, was generated that were independently and normally distributed with a mean of zero and a standard deviation of one; the desired vector of J scores $\underline{x}$ was found from $\underline{x} = L\underline{z}$. Each of the J measures in $\underline{x}$, x_j, was then transformed to a score, y_j, from a selected population with a mean (μ_j) using $y_j = x_j + \mu_j$. Nonnormal data were generated using procedures described by Fleishman (1978) and Vale and Maurelli (1983). Given a .05 level of significance, each condition was replicated 5,000 times for both power and Type 1 error rates.

Criterion for an acceptable familywise error rate. Bradley's (1978) stringent criterion was used to judge the bounds for estimates of an acceptable familywise error rate because past research, i.e., Seaman, Levin and Serlin (1991) and Keselman and Lix (1995) had indicated the potential for one or more of these P-MCPs to meet this criterion. Also, for reasons to be described when sample size is discussed, we were not as concerned with a P-MCP whose familywise α was less than Bradley's lower bound. Bradley's stringent criterion is to be considered robust when a P-MCPs empirical rate of Type 1 error ($\hat{\alpha}$) is contained in the interval $\alpha \pm 0.2\,\alpha$. For α = .05, a P-MCP was considered robust if it fell in the interval $.04 \leq \hat{\alpha} \leq .06$.

P-MCP power. Per-pair power (the probability that a true difference between two specified means will be detected) was investigated by setting two means at .3 and -.3 with the other means set at zero. Sample size (n) for each case was then found such that power was as close to .80 as possible without going below .80 (i.e., at n-1 power was less than .80). The notation n~.80 is used here to denote the latter sample size and ~.80 to denote the power for this sample size. Per-pair power was investigated because of results and reasoning given by Seaman, Levin and Serlin (1991) for fixed effects one-way designs. All pairs power (the probability that all true pairwise mean differences will be detected) was found by Seaman et al. (1991) to be highly correlated with per-pair power (r > .90), and any-pair power (the probability that at least one true pairwise mean difference

Table 1
Each Pairwise Multiple Comparison Procedure Used in Study 1, Its Abbreviation, Type I Error Similarity, Test Statistic, Critical Value α', and *q* Statistic Degrees of Freedom for Steps Between Means

Test	Letter ID	Type I Letter[c]	Test Statistic[d]	Critical Value α'	Df1[f]
Simultaneous Tests: No Omnibus Test					
(1) Tukey[a]	T	a	q	CT[e]	J[g]
(2) Keppel[b]	K	b	q	CT	J
(3) Dunn-Bonferroni	DB	c	t	$2\alpha/(J(J-1))$	_
Stepwise Tests: Preceded By Omnibus Test[h]					
(4) Schaffer-Bonferroni	SB	d	t	α/x[i]	_
(5) Fisher-Hayter	FH	d	q	CT	J-1
(6) Schaffer-Ryan-Welsch	SRW	d	q	Tukey-Welsch[j]	etc.[k]
(7) Multiple Range Ryan-Welsch	MRW	d	q	Tukey-Welsch[j]	etc.[l]
(8) Peritz	P	d	q	Tukey-Welsch[m]	etc.[l]
Stepwise Test: No Omnibus Test					
(9) Welsch	W	e	w	CT	etc.[l]

Note. When the Studentized Range Statistic, *q*, is the critical value, CON = $(2)^{-1/2}$ in Equation 5. When Student's *t* or Welsch's *w* are the critical values, CON = 1.0.
[a]Uses Equation 2 with pooled error term and degrees of freedom for error, v = (n - 1)(J - 1). [b]Called SEP1 by Maxwell (1980) to indicate use of Equation 5 with $CV_{\psi,\alpha,v} = q_{\alpha,J,n-1}$. Maxwell (1980) attributed this testing procedure to Keppel (1973). [c]Tests with the same letter have the same Type I error based on their first test. [d]The test statistics are the Studentized Range statistic *q*, Student's *t* statistic, and Welsch's *w* statistic. [e]CT (controlled by testing) indicates that the familywise level of significance (α) is controlled by the testing process and does not have to be modified by the user. [f]Df1 is the between degrees of freedom for the *q* and *w* statistics based on the number of means or number of steps between means. [g]J is the number of repeated measures. [h]The possible omnibus tests considered here were: (1) Hotelling's T^2, (2) the Greenhouse-Geisser adjusted F test, (3) The Welch-James-Johansen multivariate test statistic, (4) the Keppel Studentized Range Test. [i]Values for x are tabled in Schaffer (1986). [j]The level of significance used at each step is found as $\alpha' = \alpha_p = 1-(1-\alpha)^{p/J}$ $(2 < p < J-2)$, $\alpha_{J-1} = \alpha_J = \alpha$ this and the testing process control the familywise error rate to be α. [k]Following the overall test the next two tests of means separated by J and J-1 steps are tested using Df1 = J-1 with an additional 1 subtracted from the Df1 from a previous step at the J-2 and subsequent steps. [l]Df1 = J at the first step and 1 is subtracted from the Df1 from a previous step at the J-1 and subsequent steps. [m]The Peritz procedure makes use of the Tukey-Welsch and Newman-Keuls stepwise procedures as described by Hochberg and Tamhane (1987, pp.120-124).

will be detected) *was found to differ comparatively little among procedures, generally centering around the theoretical omnibus-test powers* (p. 581).

P-MCP sample size. We found the sample size necessary for per-pair power to be ~.80 because, based on the results of Keselman and Lix (1995), we expected these *n*'s to differ by only a few units across P-MCPs. This would be an important finding if a P-MCP failed to meet Bradley's stringent criterion only at its lower bound, but could reach power of ~.80 with only one or two more units than the n~.80 needed for a P-MCP that failed to reach Bradley's criterion at the upper bound or the n~.80 needed for a P-MCP that was much more difficult to calculate.

Results

Type I Error

As a check on our procedures, we replicated Maxwell's (1980) results for WSD, Dunn-Bonferroni, and Keppel (SEP1). We found that our results (not shown here) were consistent with Maxwell's to within $\hat{\alpha} \pm .005$. Our results when we tested the full null hypothesis (i.e., that all of the means for a given single group repeated measures design were equal) are presented in Table 2 for Wilks's overall multivariate test, WJJ, T, K, W, and DB. We included Wilks's tests as a further check on our process, because it should have found (and did find) empirical error rates that were within Bradley's stringent criteria.

Welch-James-Johansen. The results for the WJJ test indicated that with a sample size of fifteen units, the $\hat{\alpha}$'s became too liberal (i.e., $\hat{\alpha} > .06$) when the ratio of number of units to the number of measures became less than or equal to 3 to 1, i.e., $n/J \leq 3$. This result is similar to those found by Keselman, Carriere, and Lix (1993) for repeated measures main effects in unequal *n* split-plot designs. The latter authors found...*that, for normally distributed data, the number of subjects in the smallest of the unequal groups should be 2 to 3 times the number of repeated measurements minus one in order to achieve reasonable Type I error protection. (p. 311)*

Tukey and Welsch. The T and W procedures yielded very similar results. In Table 2 both procedures yielded empirical error rates within Bradley's stringent confidence bounds only when sphericity was equal to one ($\varepsilon = 1.00$). Both procedures were too liberal ($\hat{\alpha} > .06$) when sphericity was less than one, having higher $\hat{\alpha}$'s as sphericity decreased.

Keppel and Dunn-Bonferroni. In Table 2, the K procedure yielded $\hat{\alpha}$'s that became too liberal ($\hat{\alpha} > .06$) as the number of measures increased and as the measure of sphericity increased. The DB procedure yielded error rates that averaged .04, and that dropped below .04 at levels of sphericity that were close to our minimum values.

Table 2
Empirical Type I Error Rates ($\hat{\alpha}$'s) for the Full Null Hypothesis.

J	n	ε	Wilks	Welch-James Johansen	Tukey WSD	Keppel	Welsch	Dunn-Bonferroni
3	15	.51	0490	0500	0854*	0408	0788*	0356**
		.75	0532	0542	0686*	0476	0654*	0394**
		1.00	0496	0504	0496	0492	0514	0414
4	15	.40	0552	0598	0994*	0504	1028*	0382**
		.50	0482	0540	0822*	0532	0928*	0396**
		.75	0520	0542	0658*	0588	0722*	0440
		1.00	0466	0530	0460	0602*	0508	0464
5	15	.30	0462	0592	1178*	0552	1188*	0370**
		.50	0540	0662*	0980*	0606*	0948*	0404
		.75	0488	0604*	0680*	0660*	0698*	0436
		1.00	0474	0600	0460	0672*	0532	0454
6	15	.30	0508	0748*	1204*	0554	1270*	0328**
		.50	0456	0666*	0946*	0596	0970*	0352**
		.75	0590	0838*	0698*	0628*	0734*	0384**
		1.00	0494	0704*	0482	0646*	0482	0380**
8	15	.20	0520	1272*	1542*	0594	1622*	0324**
		.50	0486	1252*	1100*	0644*	1088*	0356**
		.75	0514	1262*	0762*	0676*	0764*	0380**
		1.00	0470	1168*	0458	0712*	0496	0398**
10	15	.20	0456	2092*	1852*	0730*	1940*	0398**
		.50	0544	2346*	1136*	0776*	1210*	0428
		.75	0482	2160*	0902*	0776*	0826*	0436
		1.00	0526	2212*	0542	0832*	0534	0442

Note. An * indicates that the empirical error rate was greater than Bradley's upper confidence value of .06, and an ** indicates that the empirical error rate was less than Bradley's lower confidence value of .04.

Sample Size For Power Of .80

P-MCPs not considered. As a result of the liberal $\hat{\alpha}$'s values found under normality for WJJ, T, and W, these procedures were not considered further in our sample size calculations. This caused the SB, FH, SRW, MRW, and P procedures to also be eliminated because they are dependent on the overall WJJ and K tests.

P-MCPs considered. We decided to investigate sample size for power of ~.80 for the DB procedure because it controlled $\hat{\alpha}$ below, but close to, Bradley's lower limit. We also decided to reconsider Type I error for the K procedure because its error rate seemed to be related to the unit/measure (n/J) ratio, and because the $\hat{\alpha}$'s reported in Table 2 where within Bradley's liberal criterion of robustness (i.e., $.025 < \hat{\alpha} < .075$ for $\alpha = .05$) for all values except those with J = 10 and $\varepsilon >$ .20. We considered both K's and DB's Type I error rate under both normality and nonnormality, using the sample size n~.80 for the DB procedure. This process was used because if the n~.80 needed for DB to have power of ~.80 did not control Type I error for K, the DB procedure would be a better choice.

Sample size results. The results for the latter analyses are shown in Table 3. In Table 3 the sample sizes needed for power of the DB procedure to reach ~.80 under normality are the same in most cases as the n's found under the nonnormal situation, requiring an additional unit for J=4, $\varepsilon = .40$. For these sample sizes the Type I error shown in Table 3 was similar to that found with 15 cases in Table 2 under normality, but is more conservative (approximately .02) for the nonnormal cases. The K procedure was too liberal ($\hat{\alpha} > .06$) for several cases when the n/J ratio was less than 3 and ε approached 1.0. The K procedure was conservative, with $\hat{\alpha}$ approximately equal to .04 under nonnormality.

Discussion

This study was an exploratory look at P-MCPs that had been found to control familywise Type I error in more complex designs, and therefore, were expected to also be similarly effective in the simpler single group repeated measures design. This was not found to be true. The results indicated that all of the new methods could not be recommended for use with single group repeated measures designs because their omnibus tests failed to adequately control Type I error. One reason for this may be that in the single group design the adjusted degrees of freedom (SDF) reduce to n-1 and do not involve the treatment variances as is true in more complex designs. However, a familiar and easy to calculate method, the Dunn-Bonferroni procedure, did successfully control familywise Type I error and may be recommended for use as a follow-up procedure with single group repeated measures designs.

Table 3
Sample Size (n~.80) for Power of ~.80 with the Dunn-Bonferroni Procedure and Empirical Type I Error Rates (Full Hypothesis) for DB and K Given This Sample Size.

		Normality				Nonnormality			
				Type I Error				Type I Error	
J	ε	n~.80	Power ~.80	DB	K	n~.80	Power ~.80	DB	K
3	.51	32	7748			a			
		33	8048	0314**	0366**				
	.75	8	7684			8	7782		
		9	8574	0440	0544	9	8410	0260**	0380**
	1.00	8	7476			8	7622		
		9	8402	0452	0578	9	8288	0268**	0364**
4	.40	8	6960			9	7946		
		9	8023	0392**	0566	10	8482	0280**	0414
	.50	9	7598			9	7668		
		10	8356	0432	0586	10	8222	0242**	0374**
	.75	9	7372			9	7474		
		10	8140	0480	0630*	10	8058	0212**	0350**
	1.00	9	7298			9	7432		
		10	8090	0482	0684*	10	8018	0199**	0324**
5	.30	10	7946			10	7908		
		11	8648	0348**	0556	11	8396	0368**	0368**
	.50	10	7420			10	7532		
		11	8206	0370**	0596	11	8026	0240**	0342**
	.75	10	7304			11	7932		
		11	8114	0399**	0620*	12	8402	0186**	0380**
	1.00	10	7264			11	7890		
		11	8058	0432	0680*	12	8356	0194**	0368**
6	.30	11	7744			11	7696		
		12	8448	0358**	0588	12	8316	0256**	0440
	.50	11	7568			11	7606		
		12	8314	0368**	0638*	12	8184	0240**	0380**
	.75	11	7512			11	7560		
		12	8258	0386**	0654*	12	8150	0344**	0046
	1.00	11	7512			11	7560		
		12	8258	0418	0680*	12	8150	0170**	0340**
8	.20	12	7760			12	7740		
		13	8416	0320**	0618*	13	8244	0240**	0402
	.50	12	7512			12	7522		
		13	8202	0354**	0672*	13	8088	0156**	0368**
	.75	12	7518			12	7476		
		13	8202	0374**	0696*	13	8036	0148**	0340**
	1.00	12	7470			12	7476		
		13	8148	0392**	0734*	13	8036	0142**	0330**
10	.20	13	7480			13	7548		
		14	8218	0354**	0764*	14	8046	0224**	0428
	.50	13	7410			13	7492		
		14	8148	0388**	0808*	14	8004	0172**	0384**
	.75	13	7362			14	7958		
		14	8078	0390**	0834*	15	8394	0154**	0386**
	1.00	13	7362			14	7960		
		14	8078	0406	0868*	15	8394	0132**	0370**

Note. The notation "n~.80" indicates the sample size necessary for a P-MCP to come as close to power of .80 as possible without becoming less than .80. The actual power for n~.80 is denoted by ~.80. An * indicates that the empirical error rate was greater than Bradley's upper confidence value of .06, and an ** indicates that the empirical error rate was less than Bradley's lower confidence value of .04.

[a]The variance covariance was singular under nonnormality.

BEST COPY AVAILABLE

Study 2

Introduction / Perspectives

Based on the results of Study 1, Study 2 was planned to examine the P-MCPs DB, K, and WJJ with the inclusion of the Roy-Bose simultaneous confidence intervals, R-B, (Roy and Bose, 1953) and the Studentized maximum modulus statistic recommended by Alberton and Hochberg (1984). The K and WJJ P-MCPs were included because they could prove effective in controlling FWE under conditions of nonnormality. Maxwell (1980) found the R-B P-MCPs to yield too conservative estimates of familywise error and power less than the DB procedure. This procedure was included here because Maxwell did not compare *n*'s among procedures for power at ~.80, and it was thought that the conservative R-B might be effective in controlling FWE with nonnormal data. The Studentized maximum modulus statistic (referred to here as A-H for Alberton and Hochberg) was included because it yields critical values that fall between the DB *t* statistic and the K *q* statistic. If the Studentized maximum modulus statistic proved to be successful, it could be studied as the test statistic with the SB, FH, SRW, MRW, and P procedures. Also, in Study 2 power was studied using real data which provided a wide variety of mean patterns and variance-covariance structures. This was done because past studies of one-way fixed effects designs (e.g., Klockars and Hancock, 1992; Seaman, Levin, and Serlin, 1991) had indicated that different P-MCPs were more powerful with different mean patterns. It was felt that these power differences among P-MCPs would probably be exacerbated given the different variance-covariance structures found in repeated measures designs.

Method

Data and Calculations

Data sources. One hundred real data sets described in Green and Barcikowski (1992) and in Robey and Barcikowski (1995) were used to consider the familywise error and power of the R-B, DB, A-H, and K pairwise multiple comparison procedures and the WJJ omnibus test. The primary sources of data were the *American Educational Research Journal*, the *Journal of Consulting and Clinical Psychology*, the *Journal of Speech and Hearing Research*, and *Psychophysiology*. Additionally, other studies were collected from published books, dissertations, non-published works and/or articles under submission, and paper presentations.

Effect size . The DB effect size (DB-ES) for each study was found using the equation (Barcikowski and Robey, 1985):

$$\text{DB-ES} = \sqrt{\frac{\left(\bar{Y}_i - \bar{Y}_j\right)^2}{2\left(S^2_i - S^2_j - S^2_{ij}\right)}} \qquad (6)$$

The maximum DB-ESs from each study were used for descriptive and predictive purposes in Study 2.

Power. Using the MC methods for Study 1, sample size for power ~.80 was found for each study based on the largest test parameter (TS_{ij}) of each P-MCP in the study. For example, if the A-H MCP was being considered for a given study, the study's variance-covariance matrix and means were considered to be population values. The pair of means that had the largest A-H test parameter in the population was found, and the sample size n~.80 was then found. This sample size was then used to examine control of Type I error for the R-B, DB, A-H, and K P-MCPs using the MC procedures from Study 1 to create normal and nonnormal data.

Very stringent criterion. An arbitrary decision was made (because we were not satisfied with an upper bound of .06) to use the one-tailed criterion of $\hat{\alpha} <$ .055 for determining if a P-MCP yielded adequate estimates of α. For those studies where familywise error was not controlled, (i.e., $\hat{\alpha} \geq .055$), sample size was increased until an n was found which also found $\hat{\alpha} < .055$.

Equation 5 for A-H and R-B SCI. For A-H, CON = 1 and $CV_{ij,\alpha,v} = m_{\alpha,k,n-1}$, where $m_{\alpha,k,n-1}$ is the Studentized maximum modulus statistic and Df1 = k = J(J-1)/2. Values of $m_{\alpha,k,n-1}$ were found using FORTRAN algorithms developed by Stoline, Vidmar, Sheh, and Ury (1977). For R-B, v is a value other than n-1 and is found to be v = n-J+1, Df1 = J-1, $CV_{ij,\alpha,v} = \sqrt{F_{J-1,n-J+1}}$, and CON = (n-1)(J-1) / $\sqrt{n-J+1}$.

Preliminary Analyses

Prior to considering the P-MCPs with the real data sets, the A-H procedure was considered for the data sets provided by Maxwell (1980) and for the data sets from Study 1, shown in Table 2. The results showed great promise for the A-H MCP. For Maxwell's data sets, J = 3, 4, 5; using his sample sizes of 15 and 8, the range of familywise error estimates was from .037 to .059, all within Bradley's stringent criterion. For the data sets from Study 1, the range of estimated familywise error was from .037 to .054, again, all within Bradley's stringent criterion.

Results

Descriptive Statistics From The Studies

Sphericity and repeated measures. Descriptive information for the one-hundred studies is provided in Figures 1, and 3 for values of sphericity. In Figure 1 the sphericity values follow a nearly normal distribution with a mean of .69 and a standard deviation of .20. Huynh and Feldt (1986) indicated that most values of sphericity would be greater than or equal to .75. However, in Figure 1

fifty-nine percent of the studies had values of sphericity less than .75. Figure 2 contains box-whisker plots of the values of sphericity by the number of repeated measures. In this figure it can be seen that one is more likely to obtain a value of sphericity less than .75 as the number of repeated measures increases. This is reasonable because the lower bound (1/(J-1)) of the sphericity values becomes smaller as the number of repeated measures increases. The x-axis in Figures 2 provides information on the number of studies found with a given number of repeated measures (J). The largest number of studies (43) had J = 3, 26 had J = 4, 12 had J = 5, 8 had J = 6, etc.. The largest number of repeated measures found among the studies was eleven, found in only one study.

Dunn-Bonferroni effect sizes. The DB effect sizes from the studies are shown in Figures 1 and 3. In Figure 1 the effect sizes were found to be positively skewed with a median of .72, and with 66 studies having an effect size between .01 and .98. The effect sizes found in the 34 remaining studies ranged from a very large effect size of 1.01 to a huge effect size of 7.85. The box-whisker plots in Figure 3 indicated no relationship between DB effect size and number of repeated measures.

Values of Sphericity for All Studies

```
Frequency    Stem &  Leaf

   18.00        0 *  333444444
   50.00        0 .  55555555566666666777777777
   30.00        0 *  888888999999999
    2.00        1 .  0

 Stem width:   1.00
 Each leaf:       2 case(s)
 Min            .21,    Max            1.00
 Underlined '7' is the beginning of values ≥ .75
```

Effect Size Parameters for All Studies

```
Frequency    Stem &  Leaf

   66.00        0 .  011222233334444455555666777788889
   24.00        1 .  01112333456
    3.00        2 .  3&
    7.00 Extremes    (2.8), (2.8), (3.3), (7.4), (7.9)

 Stem width:    1.0000
 Each leaf:       2 case(s)
 & denotes a single case
 Extremes in bold represent 2 cases
 Min            .0145,    Max            7.8507
```

Figure 1. Stem-and-leaf displays of sphericity and Dunn-Bonferroni effect sizes across all studies. The DB effect sizes were found for the mean differences with the largest *t* test parameter in each study.

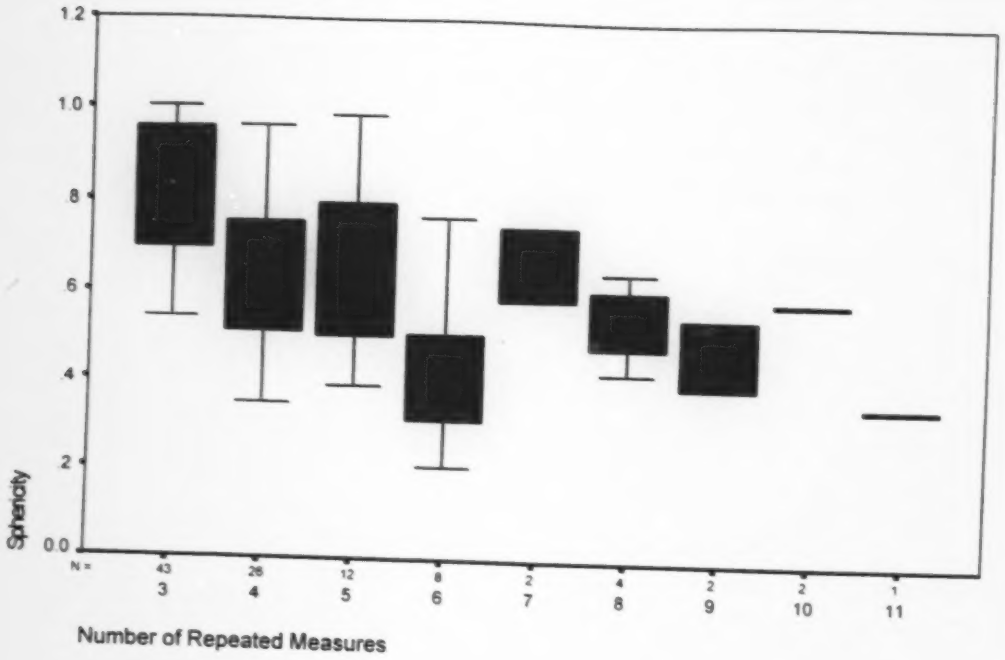

Figure 2. Box-whisker plots of sphericity by number of repeated measures across all studies.

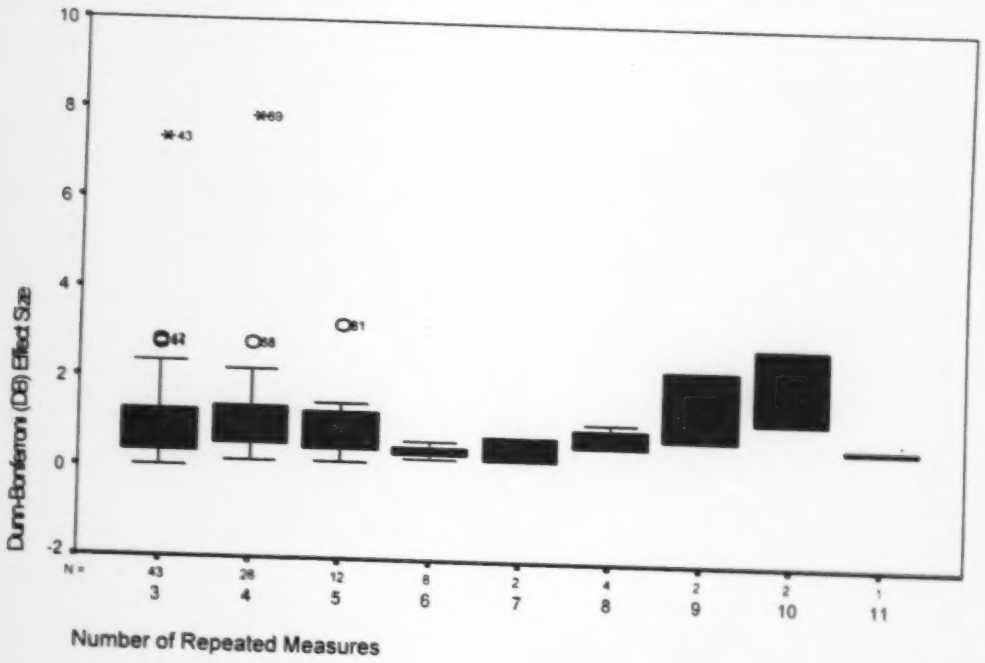

Figure 3. Box-whisker plots of Dunn-Bonferroni effect sizes (DB-ES) by number of repeated measures across all studies.

BEST COPY AVAILABLE

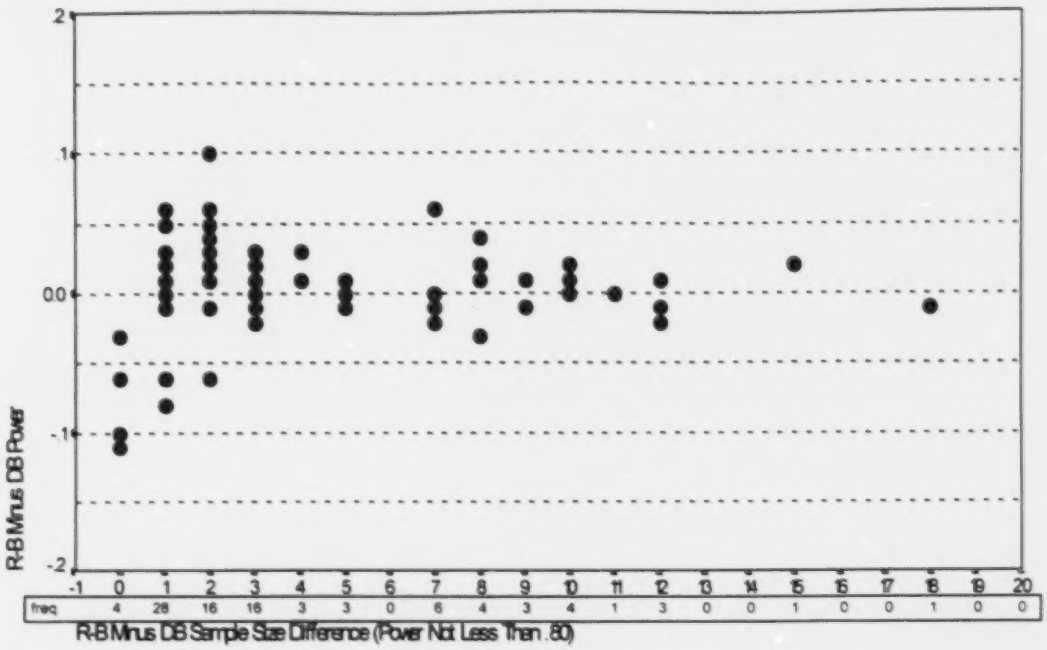

Figure 4. Scatter plot for normal data of n~.80 sample size differences by ~.80 power differences for Roy-Bose minus Dunn-Bonferroni P-MCPs. Frequencies (heights) of the sample size differences are shown below the x-axis. The coordinates of five sample size differences with values greater than 20 were deleted to improve this figure, they were: (25, -.01), (27, -.01), (34, 0.0), (38, -.01), (40, -.01). Two studies were not included because their sample sizes were so large as to make MC work difficult.

Welsch-James-Johansen (WJJ)

Results for WJJ are not given in the following sections on normal and nonnormal data because the method provided results that showed it to have poorer control of familywise error than the other methods with results slightly poorer than the K MCP. WJJ is considered in the section on prediction of familywise error.

Normal Data: Power and Sample Size Differences

Roy-Bose versus Dunn-Bonferroni. Figure 4 displays a scatter plot of the differences in the sample sizes n~.80 by the differences in the corresponding power values (~.80) for the R-B minus DB MCPs. The results indicated (as they should) that the DB MCP required either the same or smaller sample sizes in all cases, and that 64 of the cases differed by sample sizes of 3 or less; with 87 of the cases differing by 10 or less cases. The most extreme case differed by 40 sample size units with an R-B n~.80 of 377 and a DB n~.80 of 337. For those cases where the difference in sample size was zero, the power of the DB MCP was greater than the power of the R-B MCP. In Figure 4, the power differences for the four points shown at R-B minus DB = 0 sample size difference had the

following (sample sizes and R-B, DB power): (n~.80 = 12; .82, .85), (n~.80 = 9; .80, .86), (n~.80 = 6; .82, .92), (n~.80 = 6; .80, .91). Those cases with positive power differences indicated that the greater sample size required by the R-B MCP would also yield greater power. For example, at R-B minus DB = 2 in Figure 4, the largest power difference was .10 which occurred when the sample size for R-B was 6 (power = .91) and the sample size for DB was 4 (power = .81). In general, for larger sample sizes, differences in power were smaller and closer to .80. For example, consider the following largest sample sizes for each n~.80 difference ranging from 1 to 7, and their respective powers (R-B n~.80 - DB n~.80;. R-B power ~.80 - DB power ~.80): (106 - 105 = 1; .80 - .81), (69 - 67 = 2; .80 - .81), (143 - 140 = 3; .81 - .80), (64 - 60 = 4; .81 - .80), (439 - 434 = 5; .80 - .81), (67 - 60 = 6; .80 - .81). Also, in general the largest power differences came from small sample sizes. For example, the power differences shown in Figure 4 that are greater than .05 are all based on sample sizes of 10 or less.

Dunn-Bonferroni versus Alberton-Hochberg. Figure 5 displays a scatter plot of the differences in the n~.80 sample sizes by the differences in the corresponding power values (~.80) for the DB minus A-H MCPs. The n~.80 sample sizes are nearly identical for these two P-MCPs. The single largest n~.80 difference was 4 units; 63 cases had no difference, 26 cases differed by 1 unit, and 2 cases differed by 2 units. Indeed, the five units with negative differences, i.e., with smaller sample sizes for the DB procedure, represent errors in the MC procedure due to the closeness of the actual sample sizes. The ~.80 power advantage was in favor of the A-H P-MCP (as it should) when the n~.80 difference between the two procedures was zero. The largest three power differences at the n~.80 difference of 0 were -.14, -.10, and -.06, with all other differences less than -.05 (e.g., 12 differences at -.02, 15 differences at -.01, and 14 differences at .00). At the n~.80 difference of 0 there were eight ~.80 power differences at .01, favoring the DB MCP, which again represented errors in the MC procedure due to the closeness of the actual power values.

Alberton-Hochberg versus Keppel. Figure 6 displays a scatter plot of the differences in the n~.80 sample sizes by the differences in the corresponding ~.80 power values for the A-H minus K MCPs. The n~.80 sample size differences are very small for the A-H and K MCPs with 85 cases having differences between 0 and 2 and 7 cases less than a difference of 6. When the n~.80 sample size difference was 0 the power differences favored the K MCP with the single largest ~.80 power difference of -.06. Given one additional case, a n~.80 difference of 1, the A-H MCP generally had larger power values with three cases showing power differences of .09.

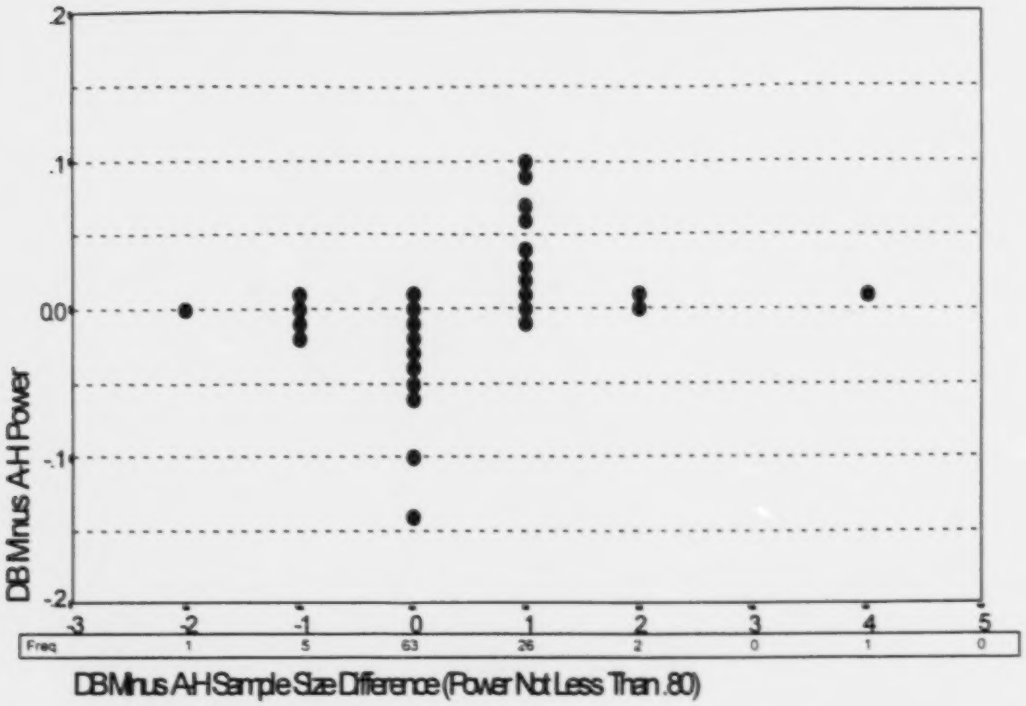

Figure 5. Scatter plot for normal data of n~.80 sample size differences by ~.80 power differences for Dunn Bonferroni minus Alberton-Hochberg P-MCPs. Two studies were not included because their sample sizes were so large as to make MC work difficult.

Sample Size with FWE < .055

In Study 1 it was observed that the K's and WJJ's Type I error was reduced with larger sample sizes. This process of considering larger sample sizes was tried with those studies whose FWE was found to be $\geq$ .055 for the A-H and K P-MCPs. For the latter studies sample size was increased until estimated FWE was less than .055. The results indicated that of the twelve A-H FWE's whose values were greater than .055, nine required increased sample sizes that were larger than n~.80 DB, two had one unit less than n~.80 DB, and one had the same size as n~.80 DB. For the K FWE forty-five studies yielded FWE $\geq$ .055. Of these, only seven had increased sample sizes that were slightly less than n~.80 DB.

Normal Data: Estimated Familywise Error

Roy-Bose. Values of familywise error were not presented because they are all known to be less than those for DB (e.g., Maxwell, 1980).

Dunn-Bonferroni. Figure 7 displays a scatter plot of the estimated familywise error by number of repeated measures based on n~.80 sample sizes for DB MCPs. The results indicated that the DB familywise errors were all less than .05 for normally distributed data.

Alberton-Hochberg. Figure 8 displays a scatter plot of the estimated familywise error by number of repeated measures based on n~.80 sample sizes for A-H MCPs. Twelve of the one-hundred cases yielded $\hat{\alpha}$'s that were greater than .055. The range of $\hat{\alpha}$'s for the other cases was between .055 and .029. On observing the 12 studies whose $\hat{\alpha}$'s were $\geq$ .055 it was found that all occurred when n~.80 / J (n to J ratio) was less than 2. However when N~.80 was increase so that $\hat{\alpha}$ < .055 for A-H, the n~.80 for DB provided a smaller sample size.

Keppel. Figure 9 displays a scatter plot of the estimated familywise error by number of repeated measures based on n~.80 sample sizes for K MCPs. Forty-five of the one-hundred cases had $\hat{\alpha}$'s that were greater than .055. The number of cases whose $\hat{\alpha}$'s were greater than .055 by number of repeated measures was: (J = 3; 17/43 or 40 %), (J = 4, 13/26 or 50%), (J = 5, 7/12 or 58%) (J = 6, 0/8 or 0%), (J > 6, 8/11 or 73%).

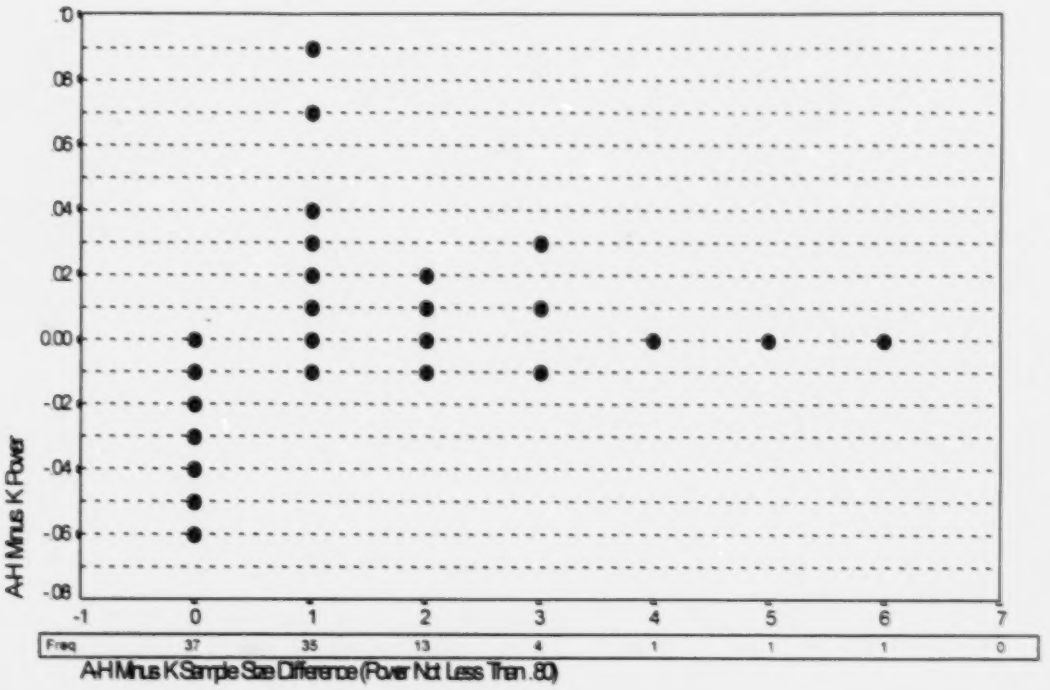

Figure 6. Scatter plot for normal data of n~.80 sample size differences by ~.80 power differences for Alberton-Hochberg minus Keppel P-MCPs. Eight studies were not included because of their sample sizes exceeded our program's limit of v = 120 for accurately computing critical values of K.

Normal Data: Prediction of Estimated Familywise Error (FWE)

Prediction using DB-ES and sphericity. In preparing the latter figures it was noticed that there appeared to be a relationship between the Dunn-Bonferroni effect size, DB-ES, sphericity and the estimated familywise errors of the testing procedures. In regressing the familywise error rates on DB-ES and

within each number of repeated measures (J) that prediction was best handled with J held constant. The results of these regressions of the familywise error rates for WWJ, K, A-H, and DB are shown in Table 4 for the values of J where the number of studies was relatively large, i.e., at J = 3, 4, 5, and 6. The multiple correlations (R) shown in Table 4 indicated that at each level of J DB-ES and ε provided very good estimates of familywise error for the WJJ, K and A-H procedures and good estimates for the DB P-MCP. For all cases the distributions of the errors of prediction were positively skewed so that most of the errors were small with the larger errors occurring with large values of FWE. For example, the maximum error for WJJ at J = 4 was .0760, but this was for a FWE of .2810 whose predicted value was .2048. In the latter case, the majority of errors were values less than the absolute value of .0070.

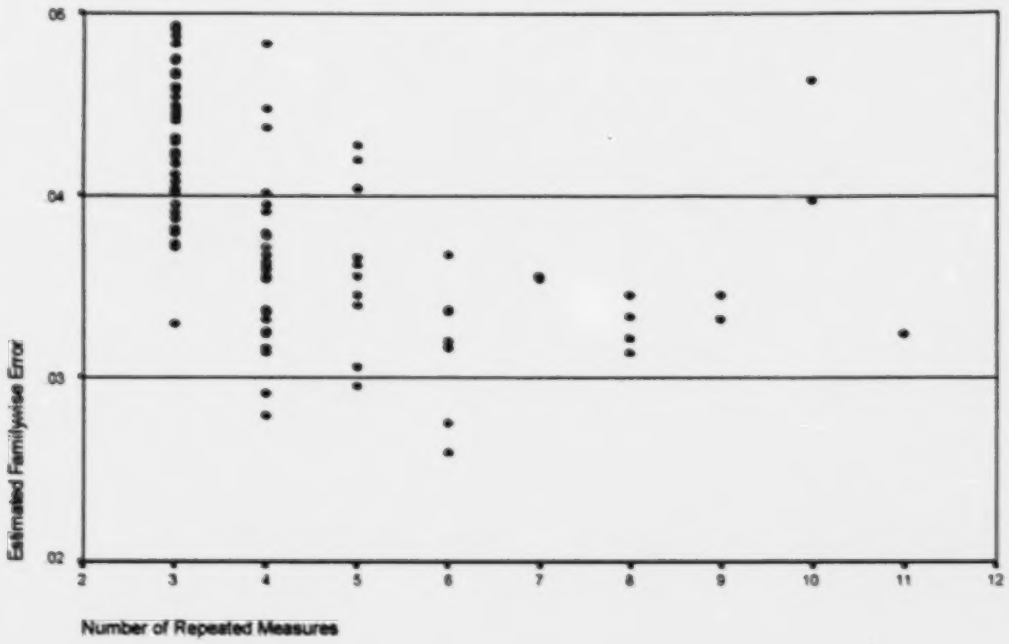

Figure 7. Scatter plot for normal data of repeated measures by estimated familywise error for the Dunn Bonferroni P-MCP, given a sample size which yields power of not less than .80.

BEST COPY AVAILABLE

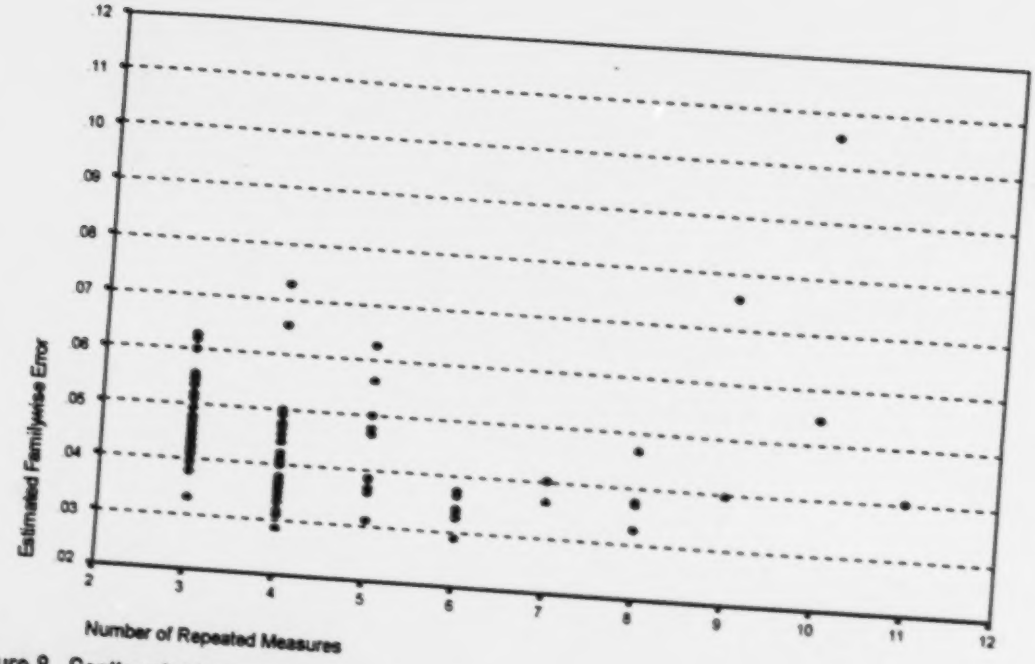

Figure 8. Scatter plot for normal data of repeated measures by estimated familywise error for the Alberton-Hochberg P-MCP, given a sample size which yields power of not less than .80.

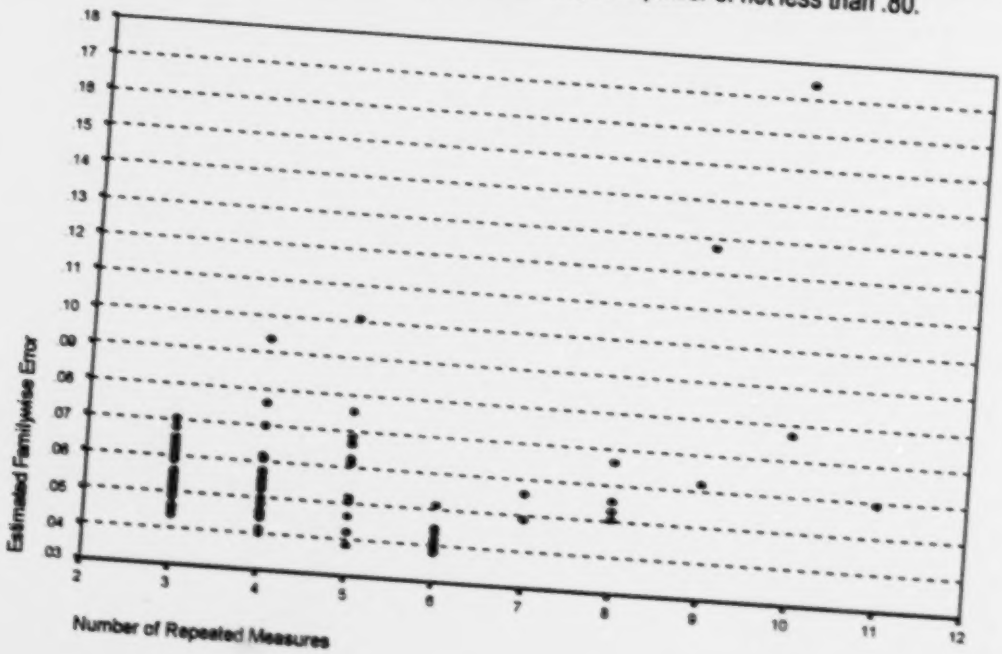

Figure 9. Scatter plot for normal data of repeated measures by estimated familywise error for the Keppel P-MCP, given a sample size which yields power of not less than .80.

BEST COPY AVAILABLE

Table 4
Regression Statistics Found in Using Dunn-Bonferroni Effect Size and Sphericity (ε) to Predict the Estimated Level of Significance for Different Paired Comparison Test Statistics, Given Different Numbers of Repeated Measures and Normally Distributed Data.

Dependent Variable Est. FWE	Regression Coefficients[a] b_0	ES	ε	R (R^2)	Std Error of Estimate	Abs. Errors Min	Max
			J = 3, n = 43				
WJJ	.043	.018	NI[b]	.981(.962)	.0045	.0004	.0130
Keppel	.036	.048	.018	.817(.667)	.0040	.0001	.0100
A-H	.023	.0047	.023	.831(.690)	.0038	.0005	.0100
DB	.024	.0017	.021	.752(.565)	.0027	.0001	.0072
			J = 4, n = 26				
WJJ	-.021	.064	.049	.976(.952)	.0220	.0008	.0760
Keppel	.030	.0064	.029	.768(.591)	.0077	.0005	.0210
A-H	.022	.0059	.022	.815(.664)	.0060	.0002	.0150
DB	.023	.0014	.018	.649(.421)	.0037	.0002	.0100
			J = 5, n = 12				
WJJ	-.069	.122	.073	.966(.934)	.0270	.0016	.0410
Keppel	.018	.024	.030	.957(.916)	.0055	.0005	.0080
A-H	.015	.013	.023	.916(.840)	.0042	.0004	.0089
DB	.019	.0039	.019	.659(.434)	.0033	.0002	.0066
			J = 6, n = 8				
WJJ	.043	.029	-.006	.926(.857)	.0019	.0001	.0023
Keppel	.025	.026	.014	.869(.756)	.0025	.0009	.0033
A-H	.032	.0089	.013	.836(.698)	.0020	.0005	.0026
DB	.033	-.014	.012	.832(.693)	.0023	.0001	.0033

Note. Estimated familywise error for each MCP was regressed on DB-ES and sphericity for each value of J, the number of repeated measures, with *n* representing the number of studies included in the regression analysis. The multiple correlation is denoted by R and the absolute values of the minimum and maximum error are shown following the standard error of estimate.
[a]The regression coefficients are unstandardized where: b_0 is the constant term, ES denotes the regression coefficient for DB-ES, and ε denotes the regression coefficient for the population measure of sphericity.
[b]NI (not in) indicates that the independent variable was not used in the regression equation.

Nonnormal Data: Power, Sample Size and Estimated Familywise Error

Sample size, power and estimated FWE. As in Study 1, the n~.80 sample size and ~.80 power values found for each study under normality were approximately the same under nonnormality. What was of interest, however, was that these same sample sizes now yielded a higher percentage of estimated FWE's that were greater than or equal to .055. The R-B and DB P-MCPs went from having no values of estimated FWE greater ≥ .055 to having 11% and 37%, respectively. The A-H and K P-MCPs increased from 12% and 45% to 41% and 65%, respectively.

Roy Bose Versus Dunn-Bonferroni. In Figure 10 is a modified stem-and-leaf display that illustrates the differences between the R-B minus DB sample sizes were the n's satisfy both the criteria of yielding power ≥ .80 and estimated FWE ≤ .055. The 30 differences that are negative indicate that the R-B MCP yielded smaller sample sizes. The DB n's in these 30 differences were increased n's (over n~.80) necessary to have FWE ≤ .055. Since some of the R-B n~.80 also had to be increased to meet the < .055 criterion, the differences indicated that these values did not have to be increased as high as the DB n's. This pattern was the same across all P-MCP sample sizes. That is, whenever a given P-MCP required that its n~.80 be increased, the increase was always in the following order: K > A-H > DB > R-B. In Table 5 are the first ten and last ten differences from the stem-and-leaf plot in Figure 10. In the first case in this table the R-B n~.80 was 6 and the DB n~.80 was 5, but these sample sizes yielded estimated FWE's of .0778 and .1086, respectively. To bring the estimated FWE to be less than .055 these two n~.80's had to be increased to 28 and 133 respectively. The last ten values in Table 5 did not have to be increased because their n~.80's yielded estimated FWE's that were all less than .055.

```
Frequency    Stem &  Leaf

     8.00 Extremes    (-105), (-86), (-78), (-67), (-63), (-56), (-53)
    22.00       -0 *  01111223334
    44.00        0 *  00111111112233345679
     3.00        1 .  1&
     4.00 Extremes    (37), (48), (50)

 Stem width not in bold:       100
 Stem width in bold:            10
 Each leaf:       2 case(s)
 & represents a single case
 Extremes in bold represent two cases
```

Figure 10. Modified stem-and-leaf display, based on statistics from nonnormal data, of sample size differences for Roy-Bose minus Dunn Bonferroni P-MCPs. Sample size was determined for power not less than .80, and familywise error less than .055.

Designs with small n's. Several of the nonnormal 30 cases where R-B provided a smaller sample size (given the criteria) than did DB, had small n to J ratios. For example, consider the three studies which required an R-B sample size of 5, 5, and 7 in Table 5, compared to the DB sample sizes of 61, 61, and 60, respectively. Although no relationship was found between RB FWE and these variables, it is interesting to note that in these designs with small n's that the conservative R-B MCP can be recommended for use.

Table 5
Examples of Studies (First 10 and Last 10 from Figure 10) with Nonnormal Data Where the Roy-Bose P-MCP Requires Smaller and Larger Sample Sizes Than the Dunn-Bonferroni P-MCP.

			n for Power ≥ .80				n for $\hat{\alpha} \leq .05$		
			R-B		DB		n		R-B - DB
J	ε	Effect Size	$\hat{\alpha}$	n~.80	$\hat{\alpha}$	n~.80	R-B	DB	Difference
Values Where R-B Requires A Smaller n									
3	.97	1.58	.0778	6	.1086	5	28	133	-105
3	.77	.42	.0640	28	.0758	27	41	127	-86
3	.88	.83	.0734	10	.0962	9	27	105	-78
5	.79	.49	.0396	33	.0830	27	33	100	-67
3	.68	.42	.0612	30	.0650	28	28	91	-63
3	.69	2.34	.0520	5	.0908	4	5	61	-56
3	.89	2.80	.0470	5	.0758	4	5	61	-56
3	.98	1.37	.0498	7	.0728	6	7	60	-53
3	.91	1.90	.0734	5	.1188	5	30	79	-49
3	.73	1.55	.0638	6	.0980	5	21	65	-44
Values Where DB Requires A Smaller n									
10	.58	2.85	.0000	14	.0166	6	14	6	8
5	.72	.38	.0176	55	.0446	46	55	46	9
8	.56	.83	.0038	25	.0490	16	25	16	9
6	.33	.45	.0114	46	.0478	35	46	35	11
9	.54	.76	.0016	30	.0320	19	30	19	11
6	.48	.41	.0106	63	.0356	51	63	51	12
3	.91	.08	.0346	785	.0414	748	785	748	37
7	.74	.24	.0044	174	.0306	137	174	137	37
5	.98	.14	.0166	382	.0338	334	382	334	48
4	.71	.14	.0232	347	.0388	297	347	297	50

Note. J = number of repeated measures, ε = population measure of sphericity, effect size (DB-ES) is for DB t, n = sample size, $\hat{\alpha}$ = the estimated familywise error rate for a given n.

Conclusions

BEST C

Tests Not Recommended

Based on the results of Study 1 (given normal data) the stepwise tests SB, FH, SRW, MRW, and P could not be recommended for use because of the failure of their possible omnibus test, WJJ, to adequately control FWE. Similarly, the T

and W procedures failed to control FWE (given normal data) and can not be recommended for use.

Tests Recommended

The results of Studies 1 and 2, given normal data, indicated that the DB P-MCP can be recommended for use with single group repeated measures data. This is because DB P-MCP was able to control FWE and because its n~.80 sample sizes were all very close to the sizes of the slightly more powerful A-H P-MCP. For nonnormal data one must take into account the power-FWE criterion that sample size should be such that power be ~.80 and that FWE should be $\leq$ .055 or .06. Based on these criteria the R-B P-MCP is recommended for use because it requires n~.80 that are generally close to the n~.80 required for DB when the DB procedure also meets the criteria, but generally requires smaller sample sizes than the DB P-MCP when the DB procedure fails to meet the criteria.

Tests Recommended Given Conditions

The A-H P-MCP can be recommended for use, given normal data, when the ratio of sample size to number of repeated measures is ≥ 2 (i.e., (n~.80 / J) ≥ 2). Given the latter condition, the n~.80 sample size of A-H was smaller than or equal to that found for DB and the A-H provides slightly better power. Given either normal or nonnormal data, there are data situations where one of the K, A-H, DB, or R-B P-MCPs best meets the power-FWE criterion in the sense of providing the smallest n~.80. For example, the K MCP required the smallest n~.80 for a nonnormal data set where the sample sizes for the MCPs were K (17), A-H (19), DB(19), R-B (30). Similarly, for another nonnormal data set, A-H provided the smallest n~.80 with $\alpha <$ .055 or .06 and sample sizes: K(21), A-H (19), DB(20), R-B (20). For DB a study was found with n~.80 of K(41), A-H (26), DB(16), R-B (25). Given normal data, one could use the regression equations provided in Table 4 to predict FWE across P-MCPs for a given n~.80, and use the power-FWE criterion to select the best P-MCP.

Monte Carlo (MC) investigations. Another approach to finding the best (smallest n~.80 which yields $\alpha <$ .055 or .06) P-MCP to use for a given repeated measures data set is to conduct a MC investigation. This approach is certainly within the scope of many investigators due to the current speeds of computers. For example, all of the results provided by Maxwell (1980) were replicated in two hours, and the MCPs considered in Study 2 could be replicated within five minutes for a single data set with an effect size $\geq$ .10 on most Pentium computers.

Recommendations for Practitioners

Recently, a large number of pairwise multiple comparison procedures were introduced to the educational research community. This study considered the use of some of the more robust of these new methods with a single group

repeated measures design over a range of nonsphericity values, given normal and nonnormal data. The results indicated that all of the new methods could not be recommended for use with single group repeated measures designs because they or their omnibus tests failed to adequately control Type I error. However, given normal data, a familiar and easy to calculate method, the Dunn-Bonferroni procedure, did successfully control familywise Type I error and may be recommended for use as a follow-up procedure with single group repeated measures designs. Also, given nonnormal data, the relatively easy to calculate Roy-Bose simultaneous confidence procedure is recommended for use in testing pairwise multiple comparisons in single group repeated measures data.

References

Alberton, Y., & Hochberg, Y. (1984). Approximations for the distribution of a maximal pairwise t in some repeated measures designs. Communications in Statistics, 13(Series A), 2847-2854.

Barcikowski, R. S. (1980). Statistical power with group mean as the unit of analysis. Final report to the National Institute of Education. Contract No. NIE-G-78-0072.

Barcikowski, R. S. & Robey, R. R. (1985, April). Sample size selection in single group repeated measures analysis. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. Psychometrika, 46, 241-255.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Collier, R. O. Jr., Baker, F. B., Mandeville, G. K., and Hayes, T. F. (1967). Estimates of test size for several test procedures on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Cornell, J. E., Young, D. M., & Bratcher, T. L. (1990). C376. An algorithm for generating covariance matrices with specified departures from sphericity. Journal of Statistical Computational Simulation, 37, 240-243.

Duncan, D. B. (1957). Multiple range tests for correlated and heteroscedastic means. Biometrics, 13, 164-176.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43, 521-532.

Green, S. & Barcikowski, R. S. (1992, April). Sphericity in the repeated measures univariate mixed-model design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association, 81, 1000-1004.

Hochberg, Y. & Tamhane, A. C. (1987). Multiple comparison procedures. New York, NY: Wiley.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. Biometrika, 67, 85-92.

Keppel, G. (1973). Design and analysis: A researcher's handbook. Englewood Cliffs, N.J.: Prentice-Hall.

Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. Journal of Educational Statistics, 19(2), 127-162.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. Psychological Bulletin, 111(1), 162-170.

Keselman, H. J. & Lix, L. M. (1995). Improved repeated measures stepwise multiple comparison procedures, Journal of Educational Statistics, 20(1), 83-99.

Keselman, H. J., Carriere K. C., & Lix L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. Journal of Educational Statistics, 18, 305-319.

Klockars, A. J. & Hancock G. R. (1992). Power of recent comparison procedures as applied to a complete set of planned orthogonal contrasts. Psychological Bulletin, 111, 505-510.

Lix, M. L. & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. Psychological Bulletin, 117(3), 547-560.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics, 5, 269-287.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Peritz, E. (1970). A note on multiple comparisons. Unpublished manuscript, Hebrew University, Israel.

Ryan T. A. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. Psychological Bulletin, 57, 318-328.

Robey R. R. & Barcikowski, R. S., (1995, April). The value of ε estimates as descriptive statistics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Satterthwaite, F. E. (1941). Synthesis of variance. Psychometrika, 6, 309-316.

Seaman, M. A. Levin, J. R. & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, 110(3), 577-586.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. Journal of Educational Statistics, 4, 14-23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Stoline, M. R., Vidmar, T., Sheh, P., & Ury, H. K. (1977). FORTRAN program: Generation of the upper α-points for the Studentized maximum modulus distribution, Western Michigan University Mathematics Report #47.

Tukey, J. W. (1953). The problem of multiple comparison procedures. Unpublished manuscript, Princeton University.

Vale, C. D. & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. Psychometrika, 48, 465-471.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. Journal of the American Statistical Association, 72, 566-575.